Proyecto de ETL y Análisis de Empresas del S&P 500: Fase 5 - Clusterización

Objetivo

El objetivo de esta fase del proyecto es que los estudiantes implementen un proceso de clusterización utilizando datos de precios históricos de las empresas del S&P 500. Los estudiantes deberán calcular métricas de volatilidad a partir de los precios diarios, y luego aplicar un modelo de clusterización para agrupar las empresas en función de estas métricas. Este proyecto les permitirá comprender cómo analizar la volatilidad de las empresas y cómo utilizar técnicas de aprendizaje no supervisado para identificar patrones en los datos financieros.

Requisitos

1. **Conocimientos previos**:
    - Python (manipulación de datos con Pandas, visualización con Matplotlib/Seaborn).
    - Conceptos básicos de ETL.
    - Conceptos de aprendizaje no supervisado, específicamente clusterización (K-means).
    - Reducción de dimensionalidad (opcional, pero recomendado).
2. **Entorno de desarrollo**:
    - **Google Colab**: Recomendado para trabajar colaborativamente y utilizar recursos en la nube sin necesidad de configuraciones locales.
    - **Cuenta en Google**: Necesaria para acceder y guardar el trabajo en Google Colab.
3. **Datasets**:
    - Archivo CSV con precios diarios de cierre de empresas del S&P 500 para un periodo determinado (3 meses). Deberán contener las columnas Date, Symbol y Close.

Instrucciones del Proyecto

Paso 1: Configuración del Entorno de Trabajo

1. **Crear un nuevo cuaderno en Google Colab**:
    - Accede a Google Colab desde tu cuenta de Google.
    - Crea un nuevo cuaderno y nómbralo "ETL y Clusterización de Empresas del S&P 500".
2. **Instalar y cargar las librerías necesarias**:
    - Asegúrate de tener instaladas y cargadas las librerías pandas, numpy, matplotlib, seaborn, scikit-learn y cualquier otra que consideres necesaria.

      !pip install pandas numpy matplotlib seaborn scikit-learn

Paso 2: Cargar y Preprocesar los Datos

1. **Cargar el dataset en el cuaderno**:
    - Sube el archivo CSV con los datos de precios diarios a Google Colab.
    - Usa pandas para leer el archivo y visualizar las primeras filas del DataFrame.
2. **Verificar y limpiar los datos**:
    - Asegúrate de que los datos no tengan valores nulos, y si los hay, decide cómo manejarlos (por ejemplo, eliminarlos o imputarlos).
    - Asegúrate de que la columna Date esté en formato datetime y ordena los datos por Symbol y Date.

Paso 3: Cálculo de Retornos Porcentuales Diarios

1. **Calcular los retornos porcentuales diarios**:

- Agrupa los datos por Symbol y calcula la variación porcentual día a día del precio de cierre (Close). Guarda estos valores en una nueva columna Return.
- Asegúrate de manejar correctamente los valores nulos que puedan surgir al calcular los retornos.

Paso 4: Cálculo de Indicadores de Volatilidad

1. **Calcular métricas de volatilidad**:
    - Para cada empresa (Symbol), calcula:
        - La desviación estándar de los retornos diarios (std).
        - El rango de los retornos diarios (diferencia entre el valor máximo y el mínimo).
        - (Opcional) La media absoluta de los retornos diarios.
2. **Crear un nuevo DataFrame**:
    - Crea un DataFrame que contenga una fila por empresa y columnas con los indicadores de volatilidad calculados.

Paso 5: Escalamiento de los Datos

1. **Escalar las métricas de volatilidad**:
    - Escala las métricas utilizando técnicas como la estandarización o normalización. Esto es importante para evitar que una métrica domine el proceso de clusterización debido a su escala.

Paso 6: Clusterización

1. **Determinar el número de clusters**:
    - Utiliza el método del codo (*Elbow Method*) o el coeficiente de silueta para determinar el número óptimo de clusters.
2. **Aplicar el algoritmo de clusterización**:

- Utiliza el algoritmo K-means para agrupar las empresas según las métricas de volatilidad.
- Asigna los clusters a cada empresa en el DataFrame.

Paso 7: Reducción de Dimensionalidad para Visualización (Opcional)

1. **Reducir la dimensionalidad**:
   - Si estás utilizando más de dos variables, aplica una técnica de reducción de dimensionalidad como PCA para reducir a dos dimensiones las características y poder visualizar los clusters en 2D.

Paso 8: Visualización de los Resultados

1. **Visualizar los clusters**:
   - Crea gráficos de dispersión para visualizar cómo se agrupan las empresas según las dos primeras métricas de volatilidad (o las componentes principales si utilizaste PCA).
   - Asegúrate de etiquetar los gráficos de manera adecuada y de incluir una barra de color que muestre a qué cluster pertenece cada empresa.

Paso 9: Análisis e Interpretación

1. **Interpretar los resultados**:
   - Analiza los clusters y describe las características de cada grupo. ¿Qué patrones observas en las empresas agrupadas? ¿Hay algún comportamiento común en las empresas dentro de un mismo cluster?
2. **Conclusiones**:
   - Escribe un breve resumen de las conclusiones obtenidas a partir del análisis y la clusterización. ¿Cómo podría este análisis ser útil para tomar decisiones financieras?

Entrega del Proyecto

- Los estudiantes deben entregar el cuaderno de Google Colab con el desarrollo completo del proyecto, incluyendo todos los pasos descritos, las visualizaciones y las conclusiones.
- Se evaluará la correcta implementación de cada paso, la claridad de las visualizaciones y la profundidad del análisis e interpretación de los resultados.